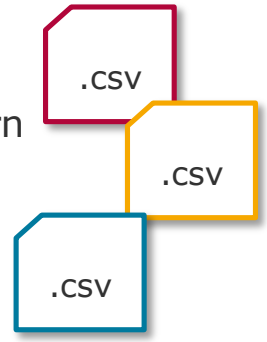
Information Integration – Task 3
Datenbereinigung

Sascha Obst, Johannes Hötter
HBRSler

09.01.2020

Kurzer Rückblick

- Uns stehen folgende Datenquellen zur Verfügung:
  - IHME: Allgemeine Daten zu Krankheiten in verschiedenen Ländern
  - GHDx: Daten zum Konsum von Tabakwaren
  - WDI:Allgemeine Bevölkerungsdaten (Einkommensschichten, ...)
- Ziel ist es, Korrelationen zwischen Einflussfaktoren auf den Konsum von Tabakwaren und der Sterblichkeitsrate aufzudecken



„Höhere Einschränkung von Werbungen zu Tabakwaren" → „geringere Anzahl an Rauchern"

„Höhere Unterstützung, mit dem Rauchen aufzuhören" → „geringere Anzahl an weibl. Rauchern"

„Höhere Steuern" → „Wesentlich weniger Jugendliche, die rauchen"

→ jeweils Betrachtung der Auswirkungen auf die allgemeine Sterblichkeitsrate

→ was sind die effektivsten Mittel, um die Sterblichkeitsrate zu verringern?

Welche Probleme bestehen noch?

- Nach der Integration der drei Quellen (IHME, GHDx, WDI) fallen noch folgende Probleme auf:
  - Uninterpretierbare Spalten
  - „Schlecht" modellierte Daten
  - Verschiedene Bezeichnungen für selben Sachverhalte, Vereinheitlichung der Bezeichnungen (sowohl von Werten als auch Spaltenbezeichnungen) praktisch
- Gut: keine Duplikate!
- Entscheidung: nur auf Views arbeiten -> besserer Überblick, um „Rohtabellen" gedanklich ausschalten zu können

Beispiele

- Bearbeiten von “schlechten” Daten

	System of National Accounts
1	Country uses the 1993 System of National Accounts methodology
2	Country uses the 1993 System of National Accounts methodology
3	Country uses the 1993 System of National Accounts methodology
4	Country uses the 2008 System of National Accounts methodology
5	Country uses the 1993 System of N
6	*null*
7	Country uses the 1993 System of N
8	Country uses the 2008 System of N
9	Country uses the 2008 System of N
10	Country uses the 2008 System of N
11	Country uses the 2008 System of N
12	Country uses the 2008 System of N
13	Country uses the 2008 System of N
14	Country uses the 1993 System of N
15	Country uses the 1993 System of N
16	Country uses the 2008 System of N
17	Country uses the 1993 System of N
18	Country uses the 1993 System of N
19	Country uses the 1993 System of N
20	Country uses the 2008 System of N
21	Country uses the 1993 System of N
22	Country uses the 2008 System of N
23	Country uses the 1993 System of N
24	Country uses the 2008 System of N
25	Country uses the 1993 System of N
26	Country uses the 1993 System of N
27	Country uses the 1993 System of N
28	Country uses the 2008 System of N

```
sql = """\
DROP VIEW IF EXISTS wdicountry_cleaned;

CREATE VIEW wdicountry_cleaned AS
SELECT "Country Code" AS "Code",
       "Short Name" AS "Entity",
       "Currency Unit",
       "Region",
       "Income Group",
       "National accounts base year",
       "National accounts reference year",
       RIGHT("SNA price valuation", 5),
       "Lending category",
       RIGHT(LEFT("System of National Accounts", 21), 4)AS "System of National Accounts Methodology (year)",
       "System of trade",
       "Government Accounting concept",
       LEFT("Latest agricultural census", 4),
       "Latest industrial data",
       "Latest trade data"
FROM wdicountry;
"""
sql = text(sql)
_ = engine.execute(sql)
```

	egory	System of National Accounts Methodology (year)
1		1993
2		1993
3		1993
4		2008
5		1993
6		*null*
7		1993
8		2008
9		2008
10		2008
11		2008
12		2008
13		2008
14		1993
15		1993
16		2008
17		1993
18		1993
19		1993
20		2008

Task 4 - Cleansing

Sascha Obst,
Johannes Hötter
09.01.2020
Chart **4**

Beispiele

- Umschlüsselungen von Werten über Mapping-Tabellen

	old_val	new_val	table_name
1	<1 year	Birth	age
2	Under 5	1 to 4	age
3	70+ years	70+	age
4	1 to 4	01 to 04	age
5	5 to 9	05 to 09	age
6	5-14 years	05 to 14	age
7	15-		
8	50-		
9	80		
10	95		
11	YLD		
12	DAL		
13	YLL		

```
# read them as a pandas file, update from old to new values
# doing this extra step to get a table which contains our mappings (as a legacy value lookup)
x_mapping = pd.read_sql_table('x_mapping', con=engine)
template = "UPDATE <table_name> SET value = '<new_value>' WHERE value = '<old_value>';"
for idx, row in x_mapping.iterrows():
    old_value, new_value, table_name = row
    sql = template.replace('<table_name>', table_name) \
        .replace('<old_value>', old_value) \
        .replace('<new_value>', new_value)
    sql = text(sql)
    engine.execute(sql)
```

Beispiele

- Erstellen von Statistiken zu Lookup-Werten (Übersichtlichkeit)

```
# statistics about lookup tables:
view_template = """\
DROP VIEW IF EXISTS <lkp_name>_occurences;

CREATE VIEW <lkp_name>_occurences AS
SELECT value, COUNT(<lkp_name>.key) FROM ihme_smoking_diseases ihme
INNER JOIN <lkp_name>
ON ihme.<lkp_name>_id = <lkp_name>.key
GROUP BY <lkp_name>.value;
"""

lkp_names = ['age', 'cause', 'measure', 'metric', 'location', 'sex']
for lkp_name in lkp_names:
    delete_sql = delete_template.replace('<lkp_name>', lkp_name)
    delete_sql = text(delete_sql)
    engine.execute(delete_sql)
    view_sql = view_template.replace('<lkp_name>', lkp_name)
    view_sql = text(view_sql)
    engine.execute(view_sql)
```

	value	count
1	01 to 04	33184
2	05 to 09	17513
3	05 to 14	17858
4	10 to 14	34094
5	10 to 24	21257
6	10 to 54	27839
7	15 to 19	37451
8	15 to 49	37986
9	20 to 24	37439
10	25 to 29	37430
11	30 to 34	37434
12	35 to 39	37424
13	40 to 44	37942
14	45 to 49	37958

Beispiele

- Löschen von nicht interpretierbaren / irrelevanten Daten

Share of women (% of women)	Share of men (% of men)	Unnamed: 5
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>	<null>	32800
<null>		
<null>		
<null>		
<null>		
<null>		
<null>		
<null>		

	Code	Year	Average cigarette price 1	Share of women... 2	Share of men...
1	MNG	2012	0	6	49
2	PHL	2012	1	9	44
3	KAZ	2012	1	8	46
4	KAZ	2014	1	7	45
5	PAK	2012	1	4	37
6	PAK	2014	1	3	37
7	KHM	2012	1	3	38
8	ZWE	2014	1	2	31
9	KHM	2014	1	2	36
10	BGD	2012	1	2	47
11	ZWE	2012	1	2	31
12	BDI	2014	2	<null>	<null>
13	CMR	2014	2	<null>	<null>
14	LBN	2012	2	28	41

Task 4 - Cleansing

Sascha Obst,
Johannes Hötter
09.01.2020
Chart **7**

Beispiele

- Pivotisierung von lückenreichen Daten

	Code	Year	Estimated daily cigarette consumption per smoker	Estimated daily cigarette consumption
1	KAZ	1998	25	25
2	UZB	1996	19	19
3	SWE	1973	<null>	<null>
4	SWE	1932	<null>	<null>
5	GRC	2008	23	23
6	JOR	2015	<null>	
7	JPN	1945	<null>	
8	LTU	2003	18	
9	MNE	1985	21	
10	BHS	2008	24	
11	CHE	1937	<null>	
12	OWID_CZS	1991	<null>	
13	ITA	1971	<null>	

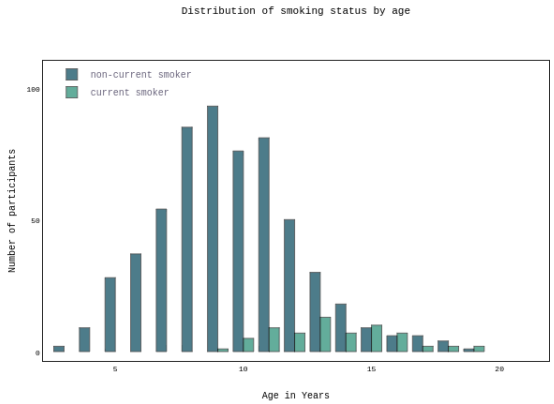
	Code	Year	Metric	Measure
1	NIC	2012	Average cigarette price	3
2	RWA	2014	Average cigarette price	2
3	MEX	2014	Average cigarette price	5
4	LKA	2014	Average cigarette price	8
5	ESP	2012	Average cigarette price	6
6	UZB	2012	Average cigarette price	2
7	JAM	2012	Average cigarette price	12
8	UZB	2014	Average cigarette price	2
9	BLR	2014	Average cigarette price	2
10	LTU	2012	Average cigarette price	5
11	ZAF	2014	Average cigarette price	5
12	NZL	2012	Average cigarette price	9
13	CHN	2012	Average cigarette price	2
14	KAZ	2014	Average cigarette price	1
15	SEN	2012	Average cigarette price	2
16	BGR	2012	Average cigarette price	6
17	DZA	2014	Average cigarette price	2
18	ITA	2012	Average cigarette price	5

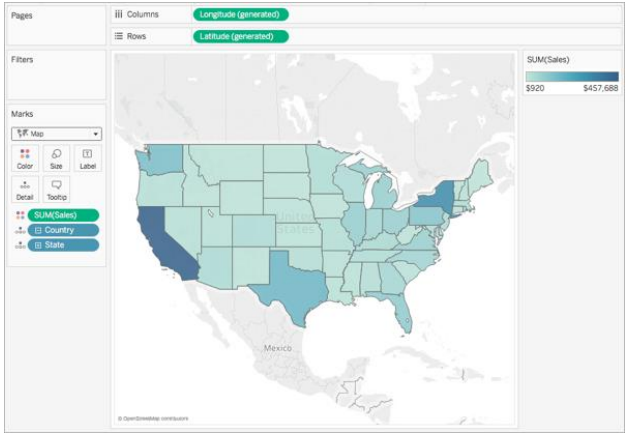
Was wir gelernt haben

- Daten können wahnsinnig schlecht modelliert sein!
  - Share of Women (in %) enthielt Werte > 30000
  - Zahlreiche Spalten „Unnamed: 5“ o.ä. in originaler Datei enthalten
  - Spalten mit langem, aber geringfügig variablem Inhalt
- Wie in vorheriger Übung: Pivotisierung oftmals hilfreich!
- Werte-Mapping Tabellen sind sehr hilfreich, um eine gute Dokumentation für Umschlüsselungen zu erhalten
  - X_Mapping
  - X_Deletion

Nächste Schritte

- Analyse der Daten auf Basis unserer Fragestellung
  - In aggregierter Form: Statistiken zu Verteilungen berechnen
  - Ggf. in Form von Karten visualisieren -> wenn, dann mit Tools wie Tableau (automatisierte Herleitung von Längen-/Breitengrad anhand ISO-3 Codes)





Task 4 - Cleansing

Sascha Obst,
Johannes Hötter
09.01.2020
Chart **10**

Danke für die Aufmerksamkeit!

Sascha Obst, Johannes Hötter
HBRSler
09.01.2020